Pierre Haessig & Marie-Anne Lefebvre – octobre 2019

Exploration de données temporelles

Cette séance de 2×1h30 a pour objectifs de vous entraîner à une activité rarement enseignée : explorer un jeu de données[1]*, en particulier temporelles. Cette activité se situe* en amont *des traitements algorithmiques (régression, prévision, identification de modèle…) qui eux sont l'objet de divers cours (statistiques, machine learning…). L'objectif est de vous donner plus d'aisance pour démarrer vos prochains travaux sur des données (étude industrielle, stage…).*

À l'issue de cette séance, vous aurez un script permettant de traiter les données qui servent de support à ce cours (questions pratiques en orange), mais aussi des notes sur quelques concepts importants en science des données (questions théoriques et mots en bleu).

Bonne pratique : toutes vos manipulations doivent être **répétables automatiquement** (e.g. après un redémarrage de Matlab). Ccl : toutes les opérations doivent *in fine* être dans un script ou une fonction.

1 Chargement & prétraitements des données

Fichier : archive météo Rennes 2015

Reconnaître et charger différents types de fichiers (CSV, largeur fixe…)

Les fichiers de données qu'on trouve « dans la nature » sont très divers (cf. Diapo schéma « Diversité des formats de fichiers de données »)

Quel est le format du fichier d'archive météo (cf. ISD-Lite 2006 format.pdf) ?

Pour traiter des séries de données, il faut les mettre dans un type de *conteneur* adapté. Pour les traitements de données en Matlab c'est `table`, et son dérivé `timetable`. (cf. Diapos). Ces types sont différents des tableaux classiques de Matlab (matrices).

Charger l'archive de météo à Rennes en 2015 dans une `table`.
Extraire la température (5e colonne, à diviser par 10)

Outils Matlab possibles : fonction `readtable` ou l'outil graphique « Import Tool[2] ». Ce dernier est pratique au premier abord, mais peu automatisable. *NB : les dates sont pour l'instant non traitées.*

Données manquantes ou aberrantes

Avant de faire des analyses compliquées, une **exploration préliminaire** permet de voir à quelles données « on a affaire ».

Tracer la série de température à Rennes (sur l'année, mais sans se préoccuper de la date).
Calculer la température moyenne (et min et max sur l'année).

Vous devriez remarquer quelques problèmes :

Y a-t-il des données aberrantes et/ou manquantes ? Si oui, combien et que faire ?

Théorie : Comment sont représentées les données manquantes[3] sous Matlab ?

Remplacer les valeurs aberrantes par des valeurs manquantes (on peut utiliser le logical indexing[4]) et refaire le tracé et calculer la moyenne avec une fonction adaptée.

Observez : lorsqu'une longue série de données contient quelques valeurs manquantes, ça ne se voit pas sur un tracé (ex : faire un zoom autour de la 2200^{e} mesure) !

NB : pour l'éventuel rebouchage des trous, voir plus bas § Rééchantillonnage et agrégation.

Dates : parsing (décodage) et représentation

Chaque fichier texte contient des dates représentées d'une façon particulière : '2018-10-22' ou bien '22/10/2018' (idem pour les heures). Matlab doit décoder (« *parser* ») ces chaînes de caractères pour en extraire le sens et stocker en mémoire l'information dans son format interne `datetime`[5] qui permet de faire des calculs avec (ex : durée entre deux dates).

Une fois la date décodée, le conteneur `table` peut être converti en `timetable`[6] qui facilite le traitement des données horodatées. Vous pouvez en particulier :

Extraire et tracer la température à Rennes sur le mois de juin.
NB : Il y a 2 méthodes pour l'extraction : la plus générique avec un `timerange`, ou par indexation logique avec la colonne des mois

2 Mise en forme & synchronisation de données

Fichiers : log capteur de qualité de l'air salle 404 (et archive météo)

Empilement/désempilement

Charger le fichier log des capteurs de qualité de l'air (humidité, CO_2, COV, température intérieure et extérieure) dans un `timetable` (e.g. décoder la date et l'heure qui sont dans des colonnes différentes puis les recombiner[7]).

Que peut-on dire de la structure (*wide* vs *narrow*[8]) de ces données par rapport à l'archive météo ? Quels sont les intérêts de chacun de ces formats ?

Désempiler (unstack) les données pour tracer l'évolution de la température intérieure.
2 méthodes possibles : plusieurs indexations logiques (une pour chaque variable) ou `unstack`

Données catégorielles & Statistiques par groupes

Certaines données sont à valeur discrète. Exemples :

- Fichier météo à Rennes : Mois = 1, 2, 3… (entier)
- Log capteur : type = 'carbon dioxide', … (chaîne de caractères)
- Champs `Month`, `Day`, `Hour`… d'un tableau `datetime`

Quel est l'intérêt de convertir la colonne 'type' en variable `categorical` ? (utiliser `whos`)

Les données catégorielles peuvent servir de Grouping Variable, par exemple pour faire des statistiques par groupes[9] ou des boxplots.

Statistique de groupe : Calculer la température moyenne/min/max pour chaque heure en salle 404. Les tracer.

Tracer un `boxplot` des températures mensuelles.

Rééchantillonnage et agrégation

Problématique : certaines analyses, comme une régression statistique, nécessitent des données *synchronisées*. Un `timetable` peut être *rééchantillonné* (avec un pas plus rapide ou plus lent) avec `retime`.

Choix de la méthode de rééchantillonnage. On peut distinguer deux cas :

- Lorsqu'on veut suréchantillonner (ou rendre périodique des mesures irrégulières), il faut choisir une méthode pour *créer des valeurs qui n'existent pas* : blocage d'ordre zéro, interpolation linéaire, filtrage médian[10]… ?
- lorsqu'on veut sous-échantillonner des données, on procède souvent à un *filtrage* ou une *agrégation statistique*. Il faut alors choisir la méthode d'agrégation (par ex. un moyennage ou une somme).

Par ailleurs, la méthode de rééchantillonnage est aussi conditionnée par :

- le type de données (variables à valeurs continues *vs* discrètes[11])
- la « largeur des trous » de données manquantes (1 heure *vs* 1 mois manquant)

Quels sont les pas d'échantillonnage des différentes données de qualité de l'air ?
Les échantillonnages sont-ils réguliers ?

Le rééchantillonnage peut servir à synchroniser les mesures des différents canaux du capteur de qualité de l'air :

Tracer un nuage de point de la concentration en CO_2 vs température intérieure.

Il peut aussi servir à combiner les données de qualité de l'air et l'archive météo (beaucoup moins de mesures).

Question ouverte : préfère-t-on agréger les données rapides ou bien interpoler les données lentes ?

Tracer un nuage de point de la température extérieure (issue du capteur de la salle 404) vs la température à Rennes de l'archive météo.

3 Recap final

Fichiers : fiche de présence salle 404 (et log qualité de l'air, archive météo)

Charger la fiche de présence. Notez que ces données sont de nature différente que les précédentes (*catégorielle* vs *quantitative*). Quelle conséquence lors d'un suréchantillonnage ?

Après resynchronisation des données, faire un nuage de point (ou un box plot) de la concentration en CO_2 (ou sa variation ?) vs présence dans la salle.

Bonus : nuage de point de la concentration en CO_2 vs température intérieure, *coloré par la présence* avec `scatter` or `gscatter`

Références

Fichiers

1) Archive de la station météo Rennes St-Jacques pour 2015. Fichier `meteo_rennes_2015.txt`
Source : NOAA Integrated Surface Database (ISD) https://www.ncdc.noaa.gov/isd

2) Log du capteur de qualité de l'air intérieur en salle 404. Fichier `log-20150309-171821.csv`

3) Fiche de présence en salle 404. Fichier `fiche_presence.csv`

NB : la salle 404 est la salle de réunion de l'équipe d'Automatique du campus de Rennes.

Liens

1. Exploratory data analysis. *Wikipedia* (2018).
2. Import Tool - Import data from file - MATLAB Documentation. https://fr.mathworks.com/help/matlab/ref/importtool-app.html.
3. Missing Data in MATLAB - MATLAB Documentation. https://fr.mathworks.com/help/matlab/data_analysis/missing-data-in-matlab.html.
4. Logical indexing. *Steve on Image Processing* https://blogs.mathworks.com/steve/2008/01/28/logical-indexing/.
5. datetime - Arrays that represent points in time - MATLAB Documentation. http://fr.mathworks.com/help/matlab/ref/datetime.html.
6. Timetables - MATLAB Documentation. https://fr.mathworks.com/help/matlab/timetables.html.
7. Combine Date and Time from Separate Variables - MATLAB & Simulink - MathWorks France. https://fr.mathworks.com/help/matlab/matlab_prog/combine-date-and-time-from-separate-variables.html.
8. Wide and narrow data. *Wikipedia* (2017).
9. Summary Statistics Grouped by Category - MATLAB Documentation. https://fr.mathworks.com/help/stats/summary-statistics-grouped-by-category.html.
10. medfilt1 - 1-D median filtering - Signal Processing Toolbox - MATLAB Documentation. https://fr.mathworks.com/help/signal/ref/medfilt1.html.
11. Retime and Synchronize Timetable Variables Using Different Methods - MATLAB Documentation. https://fr.mathworks.com/help/matlab/matlab_prog/retime-and-synchronize-timetables-using-different-methods.html.